DOCUMENT RESUME

ED 346 164 TM 018 581

AUTHOR Reinhardt, Brian M.
TITLE Estimating Result Replicability Using Double Cross-Validation and Bootstrap Methods.
PUB DATE Apr 92
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Computer Uses in Education; *Estimation (Mathematics); Heuristics; Mathematical Models; *Research Methodology; Social Science Research; *Statistical Significance
IDENTIFIERS *Bootstrap Methods; *Cross Validation; Empirical Research; Null Hypothesis; Research Replication; Statistical Package for the Social Sciences

ABSTRACT

Statistical significance is often inappropriately equated with evaluating result importance and evaluating result replicability, even though these are three somewhat different issues. The prudent researcher must separately assess each of these elements of the "research triumvirate" by using different methods. This paper focuses on two types of empirical methods for estimating research result replicability: double cross-validation, and bootstrap procedures. A commonly available statistical computer package, the Statistical Package for the Social Sciences (SPSS-X), is used to carry out the steps required for the double cross-validation procedure, and a recently developed microcomputer program package (developed by C. E. Lunneborg, 1987) is implemented to demonstrate the bootstrap logic. Both methods are applied to a heuristic data set of observed values of three independent variables and one dependent variable for a sample of 25 subjects. It is concluded that although each procedure has some shortcomings, the advantages of using either far outweigh the disadvantages. There are 5 tables of analysis data and a 19-item list of references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

stable.wp3 3/20/92

ED346164

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

BRIAN M. REINHARDT

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Estimating Result Replicability Using Double Cross-Validation and Bootstrap Methods

Brian M. Reinhardt

Texas A & M University

Department of Educational Psychology

College Station, TX 77843-4225

Paper presented at the annual meeting (session #55.15) of the American Educational Research Association, San Francisco, CA, April 24, 1992.

TM018581

BEST COPY AVAILABLE

Abstract

Statistical significance is often inappropriately equated with evaluating result importance and evaluating result replicability, even though these are three somewhat different issues. The prudent researcher must separately assess each of these elements of the "research triumvirate" by using different methods. This paper focuses on two types of empirical methods for estimating research result replicability, double cross-validation and bootstrap procedures. A commonly available statistical computer package, SPSS-X, is used to carry out the steps required for the double cross-validation procedure, and a recently developed microcomputer program package (Lunneborg, 1987) is implemented to demonstrate the bootstrap logic. Both methods are applied with a heuristic data set using multiple regression analysis so that the discussion is concrete.

Estimating Result Replicability Using Double Cross-Validation and Bootstrap Methods

Many researchers in the social sciences have invested unwisely in "statistical significance testing" stock only to find that its market value continues to shrink as the limitations of significance testing are more widely understood. Carver (1978) asserted in the Harvard Educational Review that too many researcher use statistical significance testing to support "fantasies."

One of these fantasies is to equate evaluating statistically significant results with evaluating result importance or result replicability. These null hypotheses (e.g., H_o: statistical significance = result importance) must always be rejected by the careful researcher. It may happen that in a given study the results prove to be statistically significant, important (at least by the value judgment of the researcher), and replicable, but when these three descriptors are used appropriately, they are assessed using three different methods (Thompson, 1989).

To determine if results are statistically significant, one can quickly and mechanically "decide" if a given null hypothesis (e.g., $H_o:\mu_1=\mu_2$) at a specified alpha level should be rejected or fail to be rejected. But results that are not statistically significant cannot automatically be assumed to be unimportant. This hasty generalization has produced a plethora of unpublished studies that may not have been statistically significant, but may have been useful nonetheless. Moreover, this somewhat arbitrary

discrimination procedure, which is often used by journal editors, has closed off potential research avenues (Atkinson, Furlong, & Wampold, 1982; Greenwald, 1975).

When researchers do not "achieve" statistical significance with their results, they may find it useful to ask, "At what larger n would these results be statistically significant?," since sample size is the primary influence on statistical significance (Thompson, 1989). Researchers should consider effect size in order to further evaluate the importance of their results. In multiple regression, the squared multiple correlation coefficient, R^2, is the effect size. This indicates the percent of variance of the dependent variable explained by the predictor variables. One possible effect size measure used in ANOVA is called eta squared or the correlation ratio. It indicates the percent of variance of the dependent variable that is explained by a given treatment or group. Many other effect size estimates are available in determining result importance.

Even if results are statistically significant and yield a very large effect size, they still may not be important, at least to some researchers. Result importance is inherently an inescapable personal value judgment. Mathematical calculations may help to inform these judgments, but cannot automate the value judgment process. Therefore, result importance is "judged" by carefully weighing the above mentioned factors and by considering the phenomenon being explained. Only then can the researcher make an informed value judgment as to the overall "significance"

or importance of the results.

Statistical significance testing is easy to carry out. Result importance can be rather painlessly determined by looking at several factors. But how does one determine result replicability, the essential element of the "research triumvirate"? The best way of predicting result replicability, or stability across samples, is to validate result stability empirically by conducting replications on as many independent samples as possible and by then comparing the results. Kerlinger (1986) explained the importance of replication:

> If a study is replicated and the same or similar results are found, our trust and confidence in the results are increased. If the study is again replicated and the same results are obtained, our trust and confidence are greatly increased because the probability of obtaining the same results three times by chance is lower than the probability of obtaining the same results twice. (p. 124)

In the social sciences it is often impractical to conduct numerous replication studies to determine result generalizability; instead, the stability across samples can be estimated using one of three types of techniques: double cross-validation procedure (Mitchell & Klimoski, 1986; Mosier, 1951; Pedhazur, 1982; Rowell, 1991; Thorndike, 1978), jackknife method (Crask & Perreault, 1977; Tukey, 1958), or bootstrap applications (Diaconis & Efron, 1983; Lunneborg, 1987; Thompson & Melancon, 1990).

Tukey's (1958) jackknife technique, named after the versatile and useful Boy Scout's jackknife, involves the systematic deletion of different observations or subsets of observations followed by the computation and comparison of calculated estimators (e.g., β-weight coefficients, discriminant function coefficients) derived from these revised samples. Unlike some invariance methods, the jackknife technique allows for coefficient stability to be determined using a very small sample size (Crask & Perreault, 1977). But this approach tends to focus on the influence of outliers on potential result replicability.

This paper focuses on two other techniques for estimating result stability, the double cross-validation and bootstrap methods. Cross-validation methods involve randomly dividing the original sample into subsets, conducting separate analyses, and then empirically comparing the results (Thompson, 1989). Bootstrap methods conceptually involve creating a "mega" data file by copying the original data set an enormous number of times. Random samples are then drawn from the "mega" file, analyses are conducted on each new sample, and the impacts of numerous different configuration of subjects are then compared (Crask & Perreault, 1977).

Double Cross-Validation Method

The name cross-validation is used because this procedure was originally devised to determine the validity of scoring keys in which different weights were given to the items of a test or an

inventory (Thorndike, 1978). The double cross-validation procedure (Mosier, 1951) is one of several cross-validation strategies used for splitting an original sample, called the development sample, into two samples, and then comparing various results from the samples to determine the likelihood that the original results will replicate.

The double cross-validation procedure requires seven distinct steps, each of which can be easily accomplished by using a statistical computer package such as SPSS-X. After a description of the steps, specific concepts mentioned within the task analysis such as "shrinkage" and "invariance coefficients" will be discussed, and the advantages and disadvantages of this method will be elaborated. Steps in conducting the double cross-validation procedure include:

1. The original sample of data is randomly divided into two subsamples (i.e., subsample 1 and subsample 2) with equal or unequal sample sizes. It is usually convenient to use nearly equal subsamples that are not exactly the same size.
2. Each of the variables within the two new subsets (e.g., X_{11}, X_{12},...X_{1j} for subsample 1, where the first subscript indicates the subsample number and the second subscript tells the sequence number of the predictor variable) are converted from raw scores to z scores (i.e., standard scores with a mean of 0 and a standard deviation of 1). The conversion is made by using the mean and standard deviation of subsample 1 to standardize subsample 1 data and by using the mean and standard deviation of

subsample 2 to standardize subsample 2 data (e.g., $Z_{11} = [X_{11} - X_{11}]/SD_{X11}$ for the z scores of subsample 1 (first subscript), predictor variable 1 (second subscript)).

3. Y'_{11} values are calculated using z scores from subsample 1 and Y'_{22} values are calculated using z scores from subsample 2. These are the actual regression results for each subsample.

4. Regression analyses are conducted with each subsample's z score data set yielding two regression equations, Y'_{11} (pronounced "Y-hat") for subsample 1 data (first subscript) using ß-weights derived from subsample 1 (second subscript) and Y'_{22} for subsample 2 data using ß-weights derived from subsample 2. (Note: For X, Z, ß, and Y', the first subscript indicates which subsample data set is being referenced. For Y' only, the second subscript stands for the subsample number from which the ß-weights in that regression equation were derived. For X, Z, and ß, the second subscript tells the sequence number of the predictor variable.)

$$Y'_{11} = ß_{11}Z_{11} + ß_{12}Z_{12} + ß_{13}Z_{13} + \ldots ß_{1j}Z_{ij}$$

$$Y'_{22} = ß_{21}Z_{21} + ß_{22}Z_{22} + ß_{23}Z_{23} + \ldots ß_{2j}Z_{ij}$$

5. The ß-weights are then crossed such that z scores from subsample 1 are used in the ß-weight regression equation of subsample 2 to calculate Y'_{12} and z scores from subsample 2 are used in the ß-weight regression equation of subsample 1 to calculate Y'_{21}.

$$Y'_{12} = ß_{21}Z_{11} + ß_{22}Z_{12} + ß_{23}Z_{13} + \ldots ß_{2j}Z_{ij}$$

$$Y'_{21} = ß_{11}Z_{21} + ß_{12}Z_{22} + ß_{13}Z_{23} + \ldots ß_{1j}Z_{ij}$$

6. Invariance can be evaluated by considering the shrinkage for each group. "Shrinkage" for subsample 1 is calculated by subtracting the squared multiple correlation coefficient (R^2_{12}), which is the squared bivariate correlation of Y'_{12} values and the z score values in subsample 1, from the squared multiple correlation coefficient (R^2_{11}), which is the squared bivariate correlation of Y'_{11} values and the z score values in subsample 1.

$$SHRINKAGE_1 = R^2_{11} - R^2_{12}$$

The shrinkage for subsample 2 is similarly calculated.

$$SHRINKAGE_2 = R^2_{22} - R^2_{21}$$

7. The invariance is also evaluated by calculating two invariance coefficients. The first is determined by calculating the bivariate correlation coefficient of Y'_{11} values and Y'_{12} values. The bivariate correlation coefficient of Y'_{22} values and Y'_{21} values is the second invariance coefficient.

In the task analysis, steps six and seven are crucial for evaluating the estimated invariance or stability of the research results. One way to investigate the likelihood that results will replicate is to measure the shrinkage of the multiple correlation coefficient for each subsample. Step six explains the process of calculating shrinkage. In the double cross-validation procedure, shrinkage of the multiple correlation coefficient occurs when the ß-weights are "crossed" because the ß-weights derived from the original subsample yield the highest possible correlation between the predictor variables and the dependent variable. Put differently, Pedhazur (1982) explained:

> If one were to apply a set of weights derived from one sample to the predictor scores of another sample and then correlate these predicted scores with the observed criterion scores, the resulting R would almost always be smaller that the R obtained in the sample for which the weights were originally calculated. (p.147)

The reason for shrinkage is that in calculating the weights to obtain a maximum R, the zero-order correlations are treated as if they were error-free, which is never the case. Because of this capitalization on chance, sometimes referred to as "overfitting," the original resulting R is biased upwards.

Mitchell and Klimoski (1986) concluded that shrinkage is usually reduced when predictors are chosen based on prior theory and experience-based knowledge of predictor-criterion relationships (i.e., rational procedures requiring forethought) rather than by blind empirical selection (i.e., selected with a relatively low level of rationality). Therefore, when rational procedures for selecting predictor variables are used instead of implementing "data-snooping" or stepwise multiple regression techniques (Synder, 1991), it is likely that shrinkage will be reduced, and therefore invariance or stability will increase since there is an inverse relationship between shrinkage and invariance. In other words, the degree of stability across subsamples increases as the two shrinkage estimates approach zero.

However, shrinkage formulas work poorly with small sample

sizes (i.e., less than 30 subjects per independent variable), unfavorable ratios of sample size to predictor variables (i.e., less that 3) and low multiple correlations (i.e., less than .6) (Mitchell & Klimoski, 1986; Pedhazur, 1982). Unfortunately, these conditions are prevalent in much of social science research; therefore, shrinkage formulas may be less useful under these conditions.

Invariance can also be evaluated by calculating invariance coefficients (see step seven). The shrinkage formulas described above yield results that have no set metric (e.g., an R^2 shrinkage from .9 to .7 is not equivalent to a shrinkage from .2 to 0). However, this comparison problem is not evidenced when invariance coefficients are used since they do have a set metric ranging between -1 and +1. The closer the invariance coefficients are to one, the greater the degree of confidence the researcher has that the results are replicable (Rowell, 1991; Thompson, 1989).

The advantages of the double cross-validation method are at least fourfold. First, this method does not waste data by crossing only one set of ß-weights. By crossing both sets of weights, a more rigorous approach to validation is created (Mosier, 1951; Pedhazur, 1982). Second, readily accessible statistical packages such as SPSS-X can be used easily to run the analyses needed for this procedure. A third advantage of this technique is that it saves time and money in that the researcher does not have to conduct two separate studies to determine

invariance. Finally, in most cases, this method can be used for moderate sample sizes (i.e., at least 30 subjects per predictor variable).

There are at least four disadvantages of the double cross-validation technique. First, not unlike split-half reliability coefficients which fluctuate depending on how the data is split, invariance coefficients change when different splits of the original sample are used. Second, since the sample data under examination are usually collected all at one time for convenience, any changes due to timing would not be evidenced. Third, as is always the case in research investigations, if the sample is not representative of the target population, inaccurate conclusions may be drawn by using this method. Finally, as previously mentioned, shrinkage formulas may not work well with small sample sizes, small ratios of sample size to predictor variables, and low multiple correlations.

Bootstrap Procedures

Bootstrap methods are named after the old saying about pulling yourself up by your own bootstraps, in this case by creating many samples from only one available sample (Crask & Perreault, 1977). Thirty years ago it would have been unthinkable to use the bootstrap logic. Although the actual steps required to use bootstrapping are simple from the viewpoint of practicality, the computer is a necessary partner in the process. Several microcomputer programs now allow researchers to use these methods easily (e.g., Lunneborg, 1987).

Following a description of the required steps, advantages and disadvantages of bootstrapping will be discussed. The task analysis of the bootstrap logic includes three definitive steps.

1. The original data set for each of n subjects is copied a very large number of times (e.g., 100,000,000).
2. "Bootstrap" samples of size n are randomly selected from the "mega" file, regression analyses are conducted and the ß-weights for each sample are calculated.
3. The mean, standard deviation, and median of bootstrap trials for the ß-weight estimators are calculated, and various confidence intervals are computed also. The original ß-weight estimators are compared with the bootstrap information generated from resampling.

Three advantages of the bootstrap logic overlap with those of the double cross-validation method. Like the double cross-validation method, bootstrap procedures use all of the data and do not waste any, can be quickly implemented with easy to use microcomputer programs, and provide a savings of time and money. Moreover, a unique benefit of bootstrapping is that it does not require the assumption that standard errors in the observed values be randomly and normally distributed in order to work effectively. Often this assumption is required before statistical analysis can proceed, but as Thompson and Melancon (1990) explained:

> It seems illogical to make strong assumptions that standard errors are randomly and normally distributed, when one has

data in hand that can be employed to empirically estimate standard error. (p. 8)

Computer-intensive bootstrap methods can provide estimates for the standard errors of results by using the actual data, rather than relying on the assumption that sampling error is normally distributed, which often is not the case.

Yet another advantage of bootstrapping lies in its power. Since these methods consider many configurations of subjects in their analyses, researchers can draw hypotheses about result generalizability across many different groupings of subjects.

The disadvantages of using bootstrap methods are few. As in the double cross-validation method, the influence of time factors are not considered since one data set is used instead of two or more from different research studies. Also, one must be cautious in making generalizations from a single sample since, like all statistical procedures, bootstrapping will give misleading answers for a small percentage of the possible samples (Diaconis & Efron, 1983). Finally, these methods require fairly large sample sizes to maximize their power.

Both Methods Applied to a Heuristic Data Set

Result replicability of a readily available data set from Edwards (1985, p.57) was assessed using both the double cross-validation and bootstrap methods. Observed values of three independent variables (X_1, X_2, X_3) and one dependent variable (DV) for a sample of 25 subjects were used in the multiple regression analysis. In practice, the sample size of 25 would be

too small to apply either of the two methods with confidence; however, for illustrative purposes, both will be employed.

The double cross-validation procedure was applied to the data using two separate runs of SPSS-X. The computer program is found in Table 1. Table 2 contains the complete raw data set with the randomly assigned invariance subsample numbers (i.e., 1 or 2) for each subject so that the reader can re-create the results from this example.

Insert Tables 1 and 2 about here

After the data set was randomly split into two subsamples (step 1), the SPSS-X program converted the raw scores to z scores (step 2). The conversion was made by using the mean and standard deviation of subsample 1 to standardize subsample 1 data and by using the mean and standard deviation of subsample 2 to standardize subsample 2 data. See Table 2 for a listing of the z scores.

Regression analyses were then conducted (step 3) with each subsample's z score data set producing two regression equations:

$$Y'_{11} = (+.342957*Z_{11}) + (+.60406*Z_{12}) + (+.188967*Z_{13})$$

$$Y'_{22} = (+.339154*Z_{21}) + (+.815982*Z_{22}) + (-.254246*Z_{23})$$

Then Y'_{11} values were calculated using z scores from subsample 1 and Y'_{22} values were calculated using z scores from subsample 2 (step 4). Table 2 presents these results. The crossing of the B-weight coefficients yielded two new regression equations

(step 5):

$$Y'_{12} = (+.339154*Z_{11}) + (+.815982*Z_{12}) + (-.254246*Z_{13})$$

$$Y'_{21} = (+.342957*Z_{21}) + (+.60406*Z_{22}) + (+.188967*Z_{23})$$

Finally, the invariance was evaluated by considering shrinkage (step 6) and invariance coefficients (step 7). Given that R^2_{11} was .76249 and R^2_{12} was .61121, $SHRINKAGE_1$ equals .76249 minus .61121 or .15128. The shrinkage of the squared multiple correlation coefficient for subsample one was 15 percent. Since R^2_{22} was .74564 and R^2_{21} was .57943, $SHRINKAGE_2$ equals .74564 minus .57943 or .16621. The shrinkage of R^2 for subsample 2 was 16 percent. Since the shrinkage is not zero, these estimators must be interpreted with some caution. However, since both results are similar they give support to stability across samples.

Both invariance coefficients suggest that the original regression equation for the full sample is an accurate predictor of the dependent variable in this sample and that the equation is fairly stable across samples. The two invariance coefficients were $r_{Y'11.Y'12}$, which was .8953, and $r_{Y'21.Y'22}$, which was .8815. Since both coefficients are approaching one, stability across samples is likely.

The bootstrap logic was applied to this data set by using a package of relatively "user-friendly" microcomputer programs (Lunneborg, 1987). The program gives prompts that ask for specific information (e.g., "How many bootstrap samples do you want?") in a step-by-step fashion. There is a publication that

accompanies the software (Lunneborg, 1987). One must enter the data set either by hand on the keyboard or by supplying an MS/DOS data file. In order to get ß-weight coefficients for the regression analyses, one must enter the data already converted into z scores. Raw scores are standardized by using the sample mean and sample standard deviation of the variable being considered (e.g., use mean of X_1 and standard deviation of X_1 to calculate Z_{X1}).

The program REGBOOT generated a series of bootstrap samples, and then calculated the ß-weights for each sample. The ß-weights for the original sample were also calculated by the REGBOOT program. The results were stored in an output file and used for other programs to calculate various descriptive statistics. Five hundred bootstrap samples were randomly selected from the "mega" file created by copying the heuristic data set many times. Table 3 lists a sampling of the ß-weights calculated from the 500 bootstrap samples.

Insert Table 3 about here

Next, the BOOTLV program individually calculated the mean, median, standard deviation (which is analogous to standard error), skewness and kurtosis of the ß-weights for each of the predictor variables. Table 4 contains these values. Finally, the BOOTCI program computed 90% confidence intervals for each ß-weight estimator value. Selected results are presented in Table

4. BOOTCI can calculate any width intervals (e.g., 95%, 90%, 75%) and can also provide intervals constructed using several different methods (e.g., normal theory, percentile method, bias corrected percentile, minimum width). Results from the BOOTCI program for this data set are found in Table 5.

Insert Tables 4 and 5 about here

Results from the bootstrap programs indicate that the three ß-weights derived from the original 25 subject sample data set are accurate predictors of the dependent variable in this sample and that the equation is fairly stable across samples. The means of the ß-weights for each predictor variable from the 500 bootstrap samples were very comparable to the ß-weights derived from the original sample (see Table 4).

Conclusions

Although result replicability is an essential part of the research triumvirate (i.e., statistical significance, result importance, result replicability), researchers often either ignore result generalizability or evaluate it in inappropriate ways. With the advent of computers, invariance techniques such as the jackknife, double cross-validation, and bootstrap methods can be quickly and easily applied to data sets to determine the confidence of result replicability. Although each of these procedures have some shortcomings, the advantages far outweigh the disadvantages. When actual replication of research studies is

not feasible, researchers should always employ one of these invariance procedures to determine result stability over different samples.

References

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance reviewer evaluation, and the scientific process: Is there a (statistically) significant relationship? Journal of Counseling Psychology, 29, 189-194.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312.

Crask, M. R., & Perreault, Jr., W. D. (1977). Validation of discriminant analysis in marketing research. Journal of Marketing Research, 14, 60-68.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. Scientific American, 248(5), 116-130.

Dorans, N. J., & Drasgow, F. (1980). A note on cross-validating prediction equations. Journal of Applied Psychology, 65, 728-730.

Edwards, A. L. (1985). Multiple regression and the analysis of variance and covariance (2nd ed.). New York: W.H. Freeman and Company.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82, 1-20.

Kerlinger, F. N. (1986). Behavioral research: A conceptual approach. New York: Holt, Rinehart and Winston.

Lunneborg, C. E. (1987). Bootstrap applications for the behavioral sciences (Vol. 1). Seattle: University of Washington.

Mitchell, T. W., Klimoski, R. J. (1986). Estimating the validity of cross-validation estimation. Journal of Applied Psychology, 71, 311-317.

Mosier, C. I. (1951). Problems and designs of cross-validation. Educational and Psychological Measurement, 11, 5-11.

Pedhazur, E. J. (1982). Multiple regression in behavioral research (2nd ed.). New York: Holt, Rinehart and Winston.

Rowell, R. K. (1991, April). Double cross-validation in multiple regression: A method of estimating the stability of results. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Synder, P. (1991). Three reasons why stepwise regression methods should not be used by researchers. Advances in Educational Research: Sustantive Findings, Methodological Developments, 1, 99-105.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

Thompson, B., & Melancon, J. G. (1990, November). Bootstrap versus statistical effect size corrections: A comparison with data from the finding embedded figures test. Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, LA.

Thorndike, R. M. (1978). Correlational procedures for research. New York: Gardner Press, Inc.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. Annals of Mathematical Statistics, 29, 614.

Table 1

SPSS-X Commands for Double Cross-Validation

Procedure using Heuristic Data Set

```
TITLE 'Regression Invariance Procedure'
DATA LIST FILE=ABC/ X1 1-2 X2 4-5 X3 7-8 DV 10-11 INV 13
IF (INV EQ 1)Z11=(X1-11.846)/3.436
IF (INV EQ 1)Z12=(X2-33.231)/7.328
IF (INV EQ 1)Z13=(X3-14.231)/2.774
IF (INV EQ 2)Z21=(X1-10.417)/2.392
IF (INV EQ 2)Z22=(X2-21)/6.742
IF (INV EQ 2)Z23=(X3-13.5)/2.78
IF (INV EQ 1)YHAT11=(.342957*Z11)+(.60406*Z12)+(.188967*Z13)
IF (INV EQ 1)YHAT12=(.339154*Z11)+(.815982*Z12)+(-.254246*Z13)
IF (INV EQ 2)YHAT21=(.342957*Z21)+(.60406*Z22)+(.188967*Z23)
IF (INV EQ 2)YHAT22=(.339154*Z21)+(.815982*Z22)+(-.254246*Z23)
VARIABLE LABELS YHAT11 'SUBSAMPLE 1 DATA USING SUBSAMPLE 1 BETAS'
                YHAT12 'SUBSAMPLE 1 DATA USING SUBSAMPLE 2 BETAS'
                YHAT21 'SUBSAMPLE 2 DATA USING SUBSAMPLE 1 BETAS'
                YHAT22 'SUBSAMPLE 2 DATA USING SUBSAMPLE 2 BETAS'
PRINT FORMATS Z11 TO YHAT22 (F8.5)
LIST VARIABLES=X1 TO YHAT22/CASES=500/FORMAT=NUMBERED
SUBTITLE 'Regression Using All Data'
REGRESSION VARIABLES=X1 TO DV/DESCRIPTIVES=ALL/
  DEPENDENT=DV/ENTER X1 X2 X3
TEMPORARY
SELECT IF (INV EQ 1)
SUBTITLE 'REGRESSION FOR SUBSAMPLE #1'
REGRESSION VARIABLES=X1 TO DV/DESCRIPTIVES=ALL/
  DEPENDENT=DV/ENTER X1 X2 X3
TEMPORARY
SELECT IF (INV EQ 2)
SUBTITLE 'REGRESSION FOR SUBSAMPLE #2'
REGRESSION VARIABLES=X1 TO DV/DESCRIPTIVES=ALL/
  DEPENDENT=DV/ENTER X1 X2 X3
TEMPORARY
SELECT IF (INV EQ 1)
CORRELATIONS VARIABLES=DV YHAT21/STATISTICS=ALL
TEMPORARY
SELECT IF (INV EQ 2)
CORRELATIONS VARIABLES=DV YHAT21/STATISTICS=ALL
SUBTITLE 'CHECK Z CALCULATIONS'
CONDESCRIPTIVE Z11 TO YHAT22
SUBTITLE 'INVARIANCE RESULTS'
CORRELATIONS VARIABLES=DV YHAT11 TO YHAT22/STATISTICS=ALL
```

Note. This program was adapted from Thompson (1989). It requires two runs. The first run uses the boldfaced commands. The second run includes all the commands listed above.

Table 2

Heuristic Data Set Raw Data, Converted Z Score Data, and Estimated Y Scores Using Double Crossed Regression Equations

	X1	X2	X3	DV	INV	Z11	Z12	Z13	Z21	Z22	Z23	YHAT11	YHAT12	YHAT21	YHAT22
1	11	38	10	15	1	-.24622	.65079	-1.52523	.	.	.	.02046	.83531	.	
2	7	42	16	18	1	-1.41036	1.19664	.63771	.	.	.	.35966	.33598	.	
3	12	38	18	17	1	.04482	.65079	1.35869	.	.	.	.66524	.20079	.	
4	13	36	15	16	1	.33586	.37787	.27722	.	.	.	.39582	.35176	.	
5	14	40	15	17	1	.62689	.92372	.27722	.	.	.	.82536	.89587	.	
6	15	32	11	14	1	.91793	-.16799	-1.16474	.	.	.	-.00676	.47038	.	
7	5	20	13	10	1	-1.99243	-1.80554	-.44376	.	.	.	-1.85783	-2.03621	.	
8	14	44	18	21	1	.62689	1.46957	1.35869	.	.	.	1.35945	1.06631	.	
9	14	34	12	17	1	.62689	.10494	-.80425	.	.	.	.12641	.50272	.	
10	10	28	16	11	1	-.53725	-.71384	.63771	.	.	.	-.49495	-.92682	.	
11	8	24	10	13	1	-1.11932	-1.25969	-1.52523	.	.	.	-1.43303	-1.01972	.	
12	16	30	16	18	1	1.20896	-.44091	.63771	.	.	.	.26879	-.11189	.	
13	15	26	15	16	1	.91793	-.98676	.27722	.	.	.	-.22887	-.56434	.	
14	14	24	12	15	2	.	.	.	1.49791	.44497	-.53957	.	.	.68055	1.01829
15	10	26	12	14	2	.	.	.	-.17433	.74162	-.53957	.	.	.28623	.68321
16	9	18	14	13	2	.	.	.	-.59239	-.44497	.17986	.	.	-.43797	-.60973
17	11	30	16	16	2	.	.	.	.24373	1.33492	.89928	.	.	1.05989	.94329
18	9	26	13	13	2	.	.	.	-.59239	.74162	-.17986	.	.	.21083	.44998
19	7	18	11	11	2	.	.	.	-1.42851	-.44497	-.89928	.	.	-.92864	-.61894
20	10	10	17	6	2	.	.	.	-.17433	-1.63156	1.25899	.	.	-.80744	-1.71055
21	9	12	8	12	2	.	.	.	-.59239	-1.33492	-1.97842	.	.	-1.38339	-.78717
22	10	32	18	18	2	.	.	.	-.17433	1.63156	1.61871	.	.	1.23166	.86065
23	10	18	14	12	2	.	.	.	-.17433	-.44497	.17986	.	.	-.29459	-.46794
24	16	20	15	16	2	.	.	.	2.33403	-.14832	.53957	.	.	.81284	.53338
25	10	18	12	15	2	.	.	.	-.17433	-.44497	-.53957	.	.	-.43054	-.28503

Table 3

Calculated Beta Weight Coefficients for the Sample of 25 Subjects and Seven of the 500 Random Resamplings of the 25 Subjects

Sample	Estimates of the ß-weight Coefficients		
0	.30103E+00	.66605E+00	.23807E-01
1	.40761E+00	.80936E+00	-.64835E-01
2	.20332E+00	.59310E+00	.52362E-01
3	.39495E+00	.55578E+00	.11935E+00
4	.36818E+00	.55590E+00	-.10182E+00
5	.21115E+00	.64848E+00	.28202E-02
....			
499	.26095E+00	.66821E+00	-.68605E-02
500	.18396E+00	.51799E+00	.81274E-01

Note. Sample 0 is the original sample of data for the 25 subjects. The results in the first row are the ß-weights for the three predictor variables presented in order: Z_{X1}, Z_{X2}, Z_{X3}. The rows that follow contain ß-weights from the random bootstrap samples.

Table 4

BOOTLV Bootstrap Results Across 500 Resamplings of 25 Subjects in Random Configurations

Statistic	First Predictor	Second Predictor	Third Predictor
ß-weights from Original 25	.30103	.66605	.023807
Mean of ß-weights from 500 Samples	.2925403	.6610973	.01918864
Standard Deviation	.1183305	.1288992	.12074520
Median of 500 Samples	.2932150	.661830	.0162600

Table 5

BOOTCI 90% Confidence Intervals from 500 Bootstrap Trials

	First Predictor	Second Predictor	Third Predictor
Estimator	.30103	.66605	.023807
Confidence Interval Method Used:			
Symmetric (Normal Theory)	.10534 to .49554	.45208 to .87712	-.17499 to .22316
Percentile Method	.10730 to .49149	.43928 to .86267	-.17820 to .23212
Bias Corrected Percentile	.12417 to .51435	.44150 to .86395	-.15397 to .26306
Minimum Width	.10093 to .47708	.44831 to .86809	-.19189 to .24087